UCSD Jacobs | Computer Science and Engineering

SAINT
Software & Systems Architecture & Integration Team

A Top-Down Approach to Energy saving in Data Centers

Energy Aware Service Oriented Architectures

Filippo Seracini

16 Nov. 2011

Outlook

- Data Centers
  - The energy problem
  - Current state of the art
- Our solution to the energy problem
  - Key benefits
  - Our research facility
- Conclusions

Why Data Centers are So Popular?

Inside a Data Center

Facebook Data Center

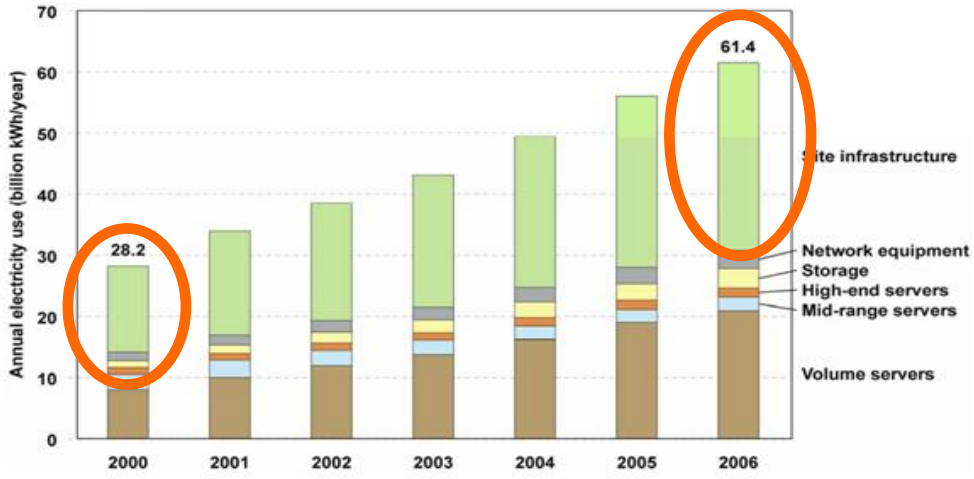
Chiller-less Container-based, Google - Belgium

Blade Servers

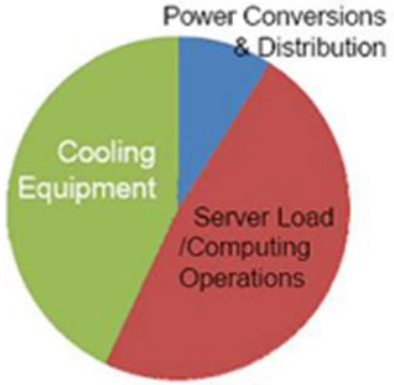
Regular and Container-based, Microsoft, Chicago

Sources:	
[i1]	[i2]
[i3]	[i4]

It's all great, but…



Electricity used by U.S. servers



- Power usage from $2B to 9$ in less than a decade for US-only data centers
- Energy consumption already up to 50% of a company's energy bill
- Already > 5% of worldwide carbon footprint

Source: [s1], [i5]

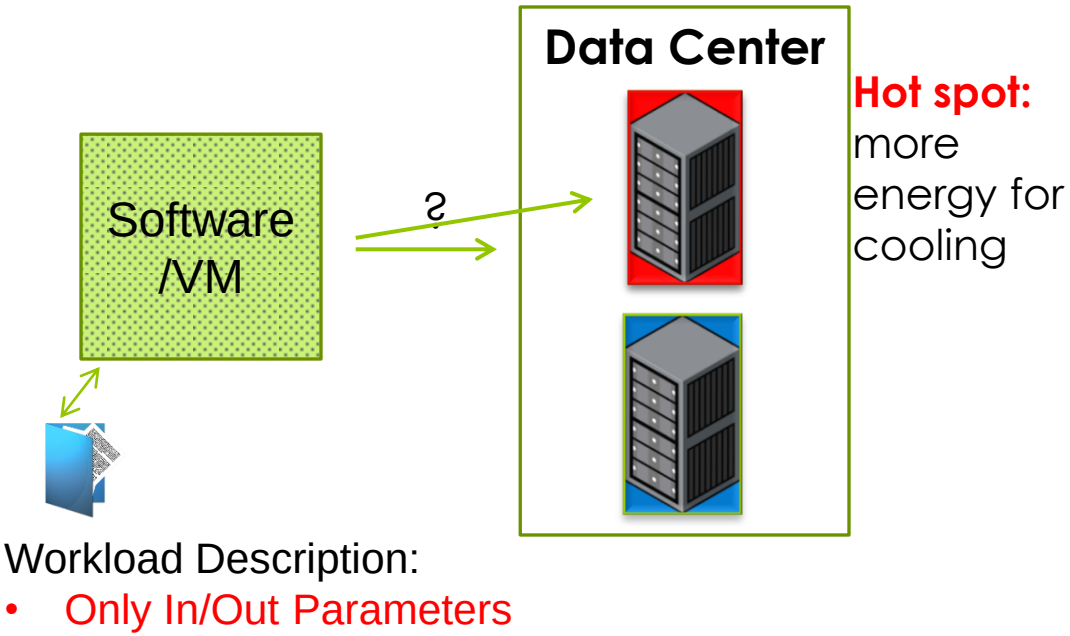
Good News: A Lot of Room for Improvement!

- Average servers' idle time 70-80% because of overprovisioning for SLAs [s2]
  - Use more information from SLAs
- At idle, server's power draw is 60% of peak power [s3]
  - Servers consolidation
- Cooling as major factor in data center design and energy efficiency
  - Avoid hot spots

State of the Art

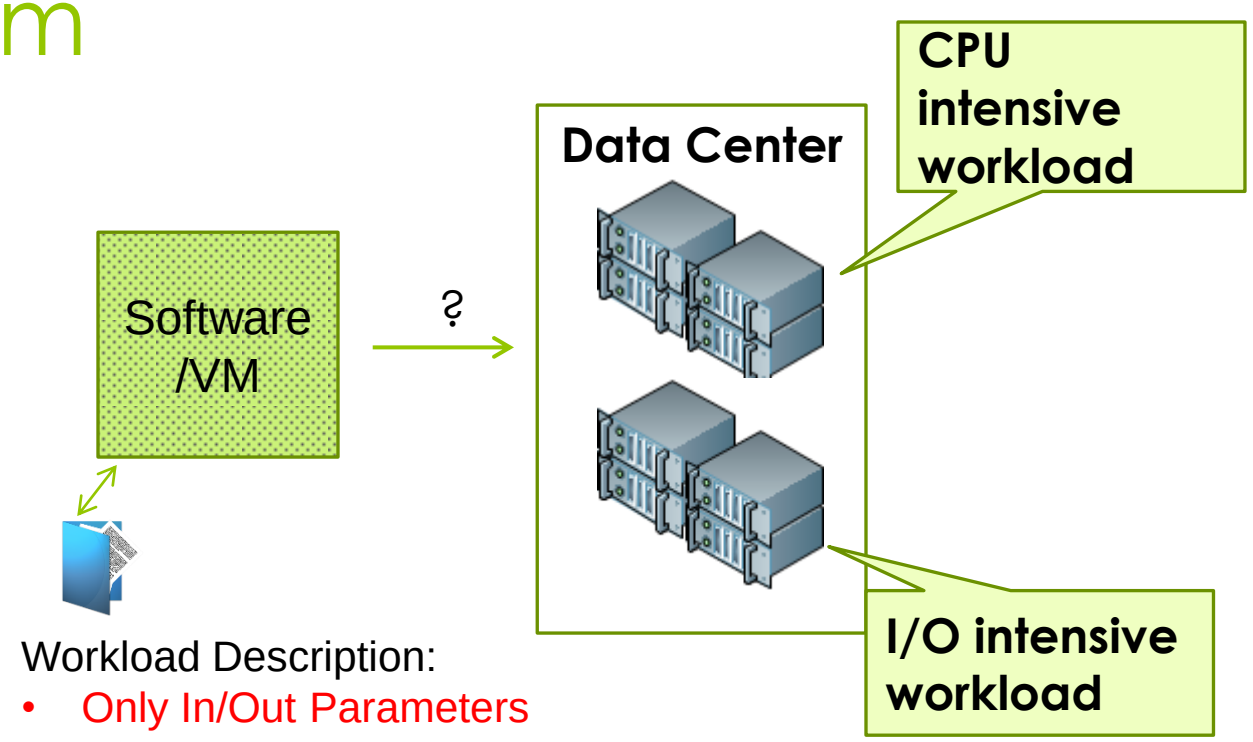
- Many existing hardware and software solutions for server consolidation
  - Software Virtualization (i.e. VMware, Xen, KVM, Hyper-V, …)
  - Hardware support (i.e. Intel VT-x, AMD-V)
- About cooling:
  - Chiller-free data centers. Move the computation when it's too hot outside.
  - Hot-cold aisle

The Cooling Problem



- Which resource should be allocated to the software/VM?
- Where is more convenient to allocate it?

The Resource Consolidation Problem



- Where should a CPU-intensive workload be allocated?
- By limiting resource contention, more workloads can be allocated on the same physical machine
- Less physical servers needed

Open questions

- How many VMs per physical machine?
- How to avoid resource contention?
- Is the SLA satisfied?
- Where to allocate the VMs?
- What kind of workload is running on a VM?

Our Solution

Have the application layer exposing more information on its functioning, so that informed decisions can be taken and the energy efficiency increased.

(Web) Services

- A service is a function, typically exposed over a network, that provides a functionality to other services or users.
- (Web) services are a dominant type of application running on today's data centers
- Examples:
  - E-commerce
  - Cell Phone apps
  - Webmail
  - Bing/Google map
  - Video streaming
  - …

Service Description Example

```
<?xml version="1.0"?>
<definitions name="expressDeliveryCompany"
        ...namespaces...>
   <message name="BookDeliveryRequest">
      <part name="bookTitle" element="string"/>
      <part name="address" element="string"/>
      <part name="deliveryTime" element="date"/>
```

Input

- Only Input-Output information
- No QoS data

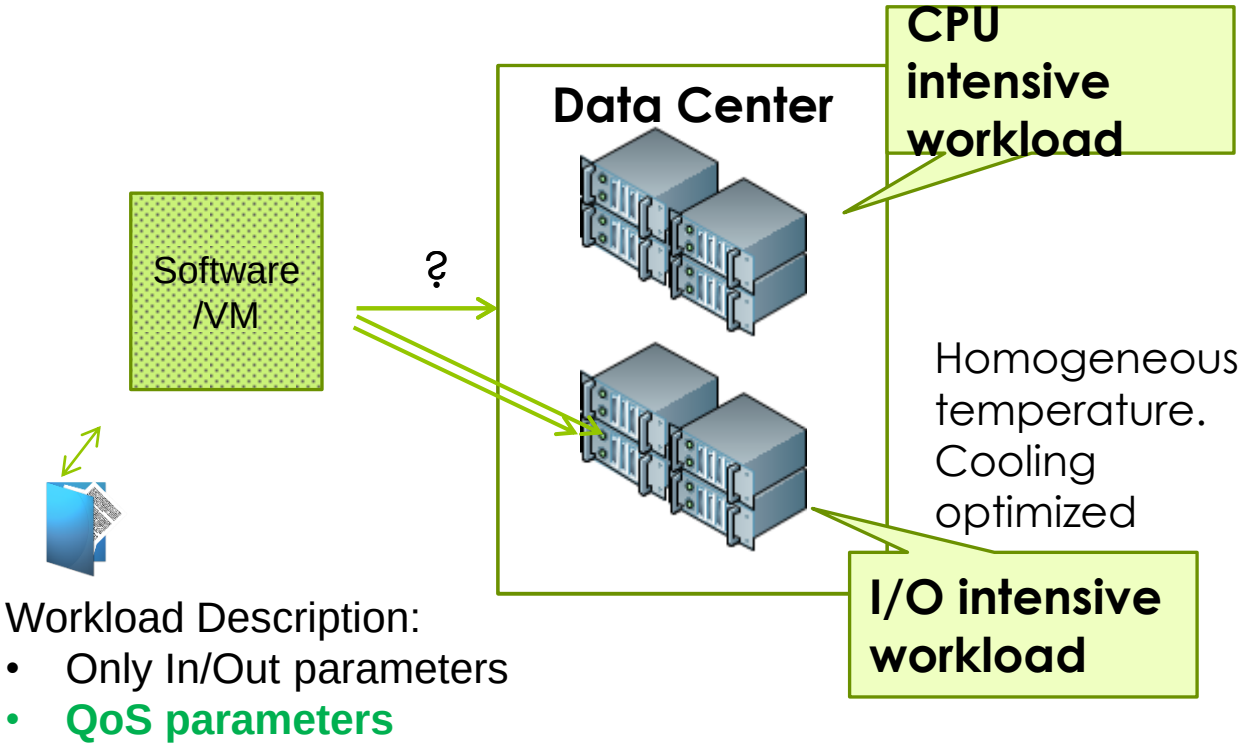
```
         <input message="BookDeliveryRequest"/>
         <output message="ConfirmSchedulingDelivery"/>
      </operation>
   </portType>
   <binding name="BookDeliverySchedulingSoapBinding" type="BookDeliverySchedulingPortType">
            …..communications and interaction protocols….
   </binding>

   <service name="BookDeliverySchedulingService">
         <port name="BookDeliverySchedulingPort" binding="tns:BookDeliverySchedulingBinding">
         <soap:address location="http://example.com/expressDeliveryCompany"/>
      </port>
   </service>
</definitions>
```

Endpoint

Key Insight

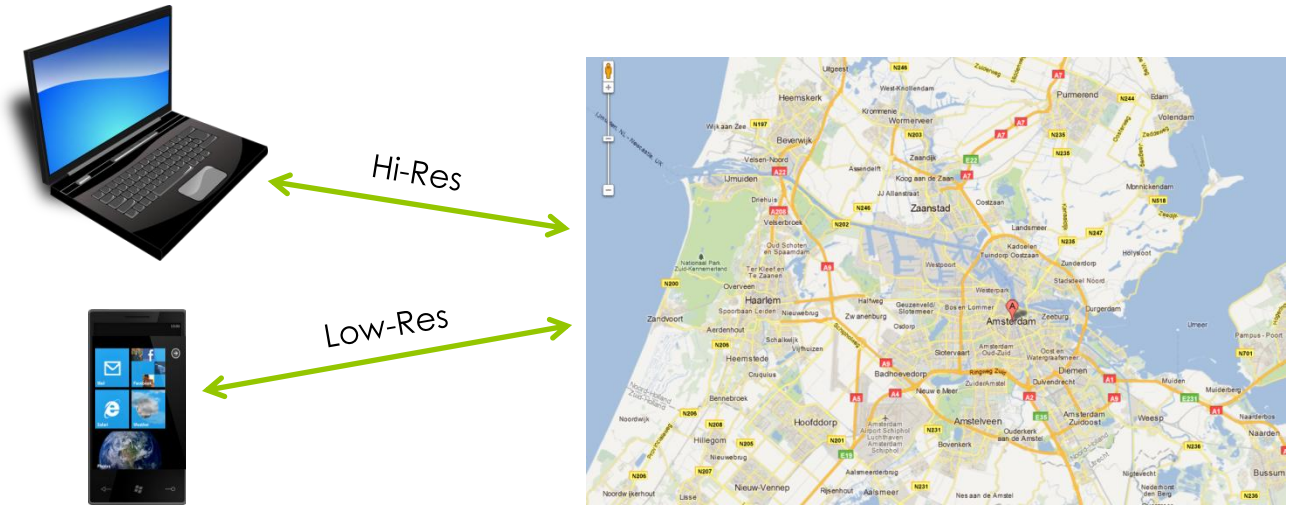
- Associate services with a richer interface specification that exposes also the non-functional parameters
- Use the additional information to:
  1. Improve workload allocation strategies
  2. At run time, enable a dynamic informed selection of the most suitable service

Allocation Time Benefits



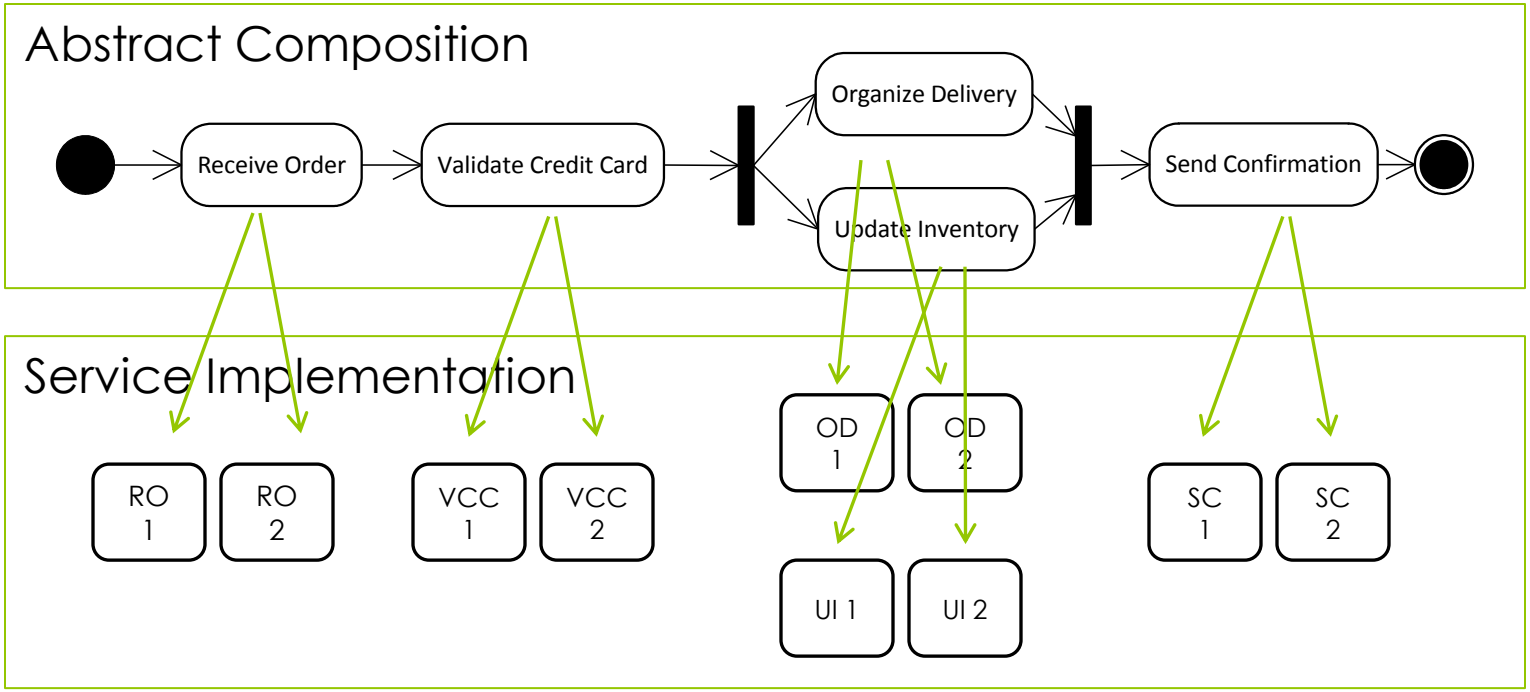
- At resource allocation time, the service/VM bundle can be deployed in the most energy efficient location
- Cooling optimization
- Maximization of resource consolidation

Run time Benefits: Service Dynamic Selection



- Create a dynamic infrastructure that dynamically selects the appropriate service to satisfy the constraints and to adapt to changes in the execution environments

Run time Benefits: Service Dynamic Composition



The system is getting overloaded. Need to reconfigure the system to keep satisfying the SLA.

Our Research Facility

- Container-based Data Center
- Instrumented with 1000+ sensors
- Variety of hardware solutions
- Real time monitoring system (http://glimpse.calit2.net/)

Summary

- Exploit the wealth of information from the application layer to:
  - Optimize cooling and server consolidation by selecting the most energy efficient location in the data center
  - Optimize run time behavior (SLAs) by dynamically selects the appropriate service
- Our approach is complementary to existing solutions

Thank you!

Questions?

References

[i1] http://www.datacenterknowledge.com/archives/2009/04/17/a-look-inside-facebooks-data-center/

[i2] http://www.datacenterknowledge.com/archives/2009/07/15/googles-chiller-less-data-center/

[i3] http://www.gis.ttu.edu/center/images/Hardware/BladeServers.jpg

[i4] http://news.cnet.com/2300-10805_3-10001679.html

[i5] http://www1.eere.energy.gov/femp/program/dc_energy_consumption.html

[s1] *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431, U.S. Environmental Protection Agency ENERGY STAR Program, 2007* .

[s2] David Meisner, Brian T. Gold, and Thomas F. Wenisch. 2009. PowerNap: eliminating server idle power. *SIGPLAN Not.* 44, 3 (March 2009), 205-216.

[s3] Luiz André Barroso and Urs Hölzle. 2007. The Case for Energy-Proportional Computing. *Computer* 40, 12 (December 2007), 33-37.